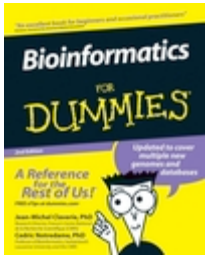
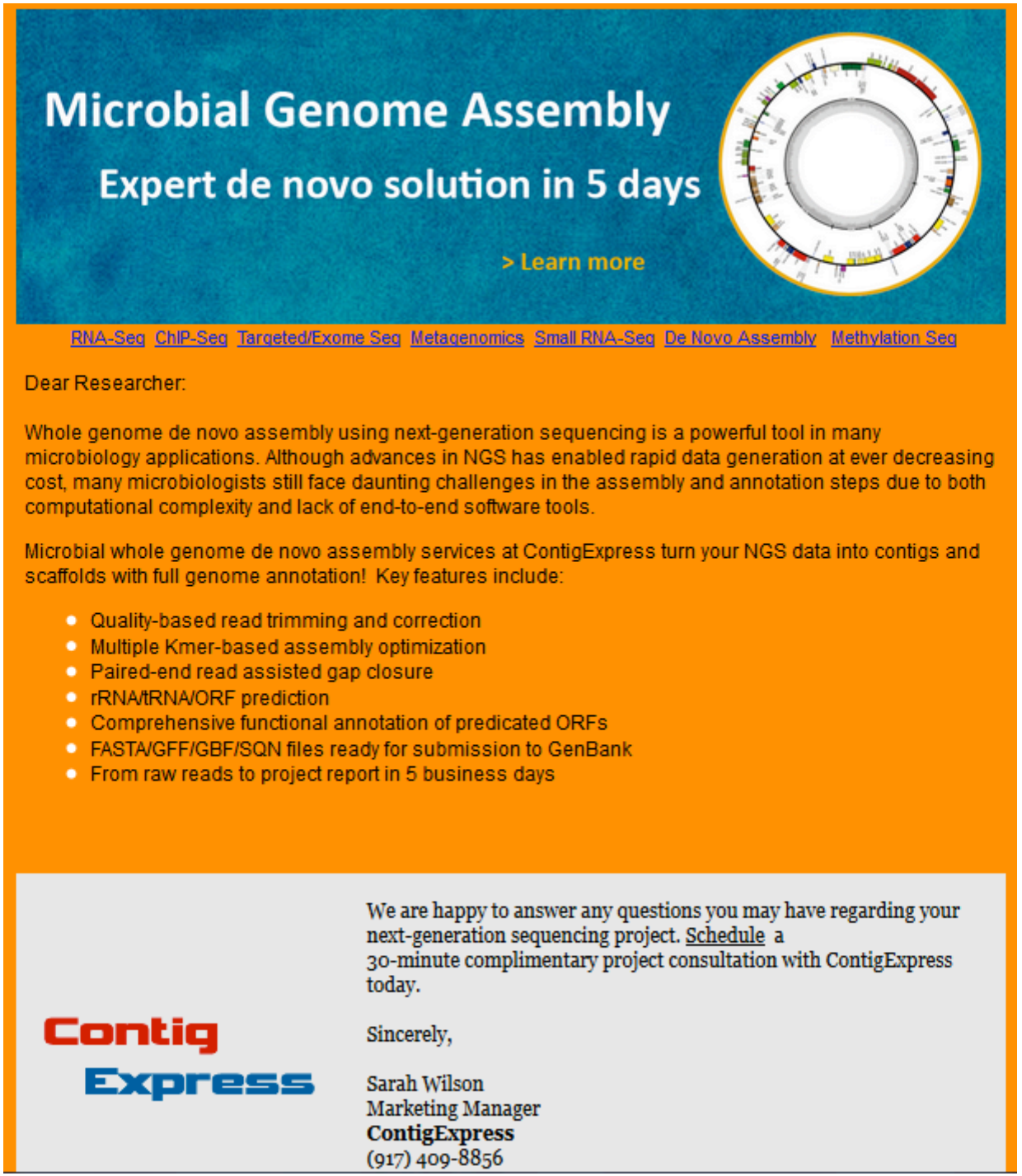


Genomics for Dummies

Bio-informatics and Comparative Genomes Analysis: Jean-Michel Claverie

6 -18 mai 2013

Do I need to know more than just this telephone number ?



Introduction

- Try to learn/understand things at the conceptual level, don't be side tracked by details
- Important biological discoveries don't need statistics, in fact many basic fact are not « statistically significant »
- **LOOK** at your data: automated processing is necessary but « automated discovery » is a myth
- Try to understand the detail of the experimental protocol
- Ultimately look for surprises, not what is known
- Know the current dogma in your field of research (for eventually destroying it)
- Know your basic biochemistry and cellular biology:
  (was is anecdotal, what is central, what does not suffer exceptions)

From DNA to the best possible paper

- Read assembly (ies !): the crazy part
- From N contigs to a single one: being lucky
- Why having a complete genome is important
- Automated genome annotation is a myth

  (in fact, genome annotation is NOT possible by Bioinformatics only)
- Quickly assess your « genome potential »
  - New Global features (ex: new genetic code)
  - New unique features (ex: no DNA polymerase, new pathway)
  - New taxonomical feature (ex: new domain of Life)
  - Incredible lack of novelty (ex: unexplained pathogenicity)

Read Assembly (ies !): the crazy part

- In general, assembly pipeline won't work as published (problems are NOT reported)
- Use different softwares, different parameters,
- Understand the software principle
- Know the default option, and what they do (use colleagues and forums)
- The true assembly is not always the best one (N50)
- MAP your read back, LOOK at it
- When confused, analysecontigs using size, coverage, GC%, BlastX, …,
- Dotplot is the greatest tool
- Tend to believe recurrent contigs

A real thing (2.5 Mb genome)

We assembled the ***xxxx*** genome using a combination of sequencing technologies: Illumina Hiseq 2000 (273,457,838 100 bp paired-end reads), 454 GS FLX+ (542,034 reads) and Pacbio RS (51,552 reads).

1) Initial assembly was done on the Illumina dataset using the Velvet assembler with stringent parameters (k=91), resulting in 62 contigs > 1kb, for a total size of 2,474,718 nt.

2) In parallel we assembled the 454 dataset using Newbler with default parameters, resulting in 58 contigs > 1kb for a total size of 2,476, 885 nt.

3) The two assemblies were merged by shearing the contigs into 2 kb sequences overlapping by 500 nt, and assembled using Phrap.

The merged assembly resulted in 20 contigs > 1 kb, for a total size of 2,479,299 nt.

4) The Pacbio dataset was then used to scaffold these contigs using the AHA hybrid assembly module from the SMRTpipe package . In addition we used the PBjelly tool with the same dataset to fill the gaps between connected contigs, resulting into a **single contig assembly**.

5) Sequencing errors were subsequently corrected by mapping the Illumina reads using Bowtie2 , and Gap5 was used to compute the consensus sequence.

We got a high-quality genomic sequence of 2,473,870 nt.

Verifying

1) We mapped the Illumina, 454 and Pacbio datasets back to the genome using respectively Bowtie2, Newbler and Blasr .

2) This allowed us to identify 12 regions with atypical sequence coverage.

3) All these regions were check by PCR, resequenced using standard Sanger technology and confirmed the genomic sequence.

Phaeocystis globosa virus, 460 kb: assembly

The raw paired-end dataset was submitted to **Velvet** with a Kmer of 75 and a mean insert size of 260 leading to 12,789 contigs **(Contigs_V75),** a n50 of 248 and a longest contig of 20,023 bp.

The raw dataset was filtered to minimize the impact of poor quality reads: i) the reads were trimmed using a quality score of 24 as a threshold, ii) the trimmed reads were removed to keep read ≥101 bp, iii) depending on the presence of their mate pair, the remaining 34,728,278 reads were split into 2 clean datasets, paired or unpaired, with a mean quality value above 32 to over 40 at each position. These datasets were submitted to **VelvetOptimiser** v2.1.7 with a Kmer varying between 85 and 99. The best result was obtained with a Kmer 95, an expected coverage of 275 and a coverage cutoff of 43.4 leading to 53 contigs **(Contigs_V95)** and n50 of 106,510. The longest contig obtained was 140,389 bp long.

Contigs_V75 and Contigs_V95 were submitted to **Phrap** with a minimum length of matching word set to 20 bp and a minimum scoring alignment set to 40. **FOUR** contigs were identified as PgV sequences with respective lengths of 4792, 62484, 174038 and 218837 bp. These contigs were then ordered and connected at once using **PCR** and Sanger sequencing of the amplicons.

Phaeocystis globosa virus, 460 kb: interpreting

All single reads where then mapped back without error to this assembly and mean coverage was computed for each contig and plotted. Three contigs exhibited a very large coverage (>4500) and a similar low (G+C) content ≤35%. They were interpreted as follows:

1) Contig #4130 (4,739 bp) and contig #4132 (13,908 bp) with a coverage≅ 33,000 and 35% (G+C). Gap closing between these contigs was achieved by PCR and bioinformatics methods (see below). This provided the PgVV virophage genome sequence;

2) Contig #4133 (459,492 bp) with a coverage ≅ 4500 and 32% (G+C). This corresponds to the final deposited PgV-16T sequence.

Another large contig (103,459 bp, contig #4125) exhibited a much lower coverage (≅ 500) and 35% (G+C). This contig sequence was found to be very similar (85% identical at the nucleotide level) to the published chloroplast genome sequence of *Phaeocystis Antarctica* (accession# NC_016703). This sequence most likely originated from *P. globosa* chloroplast DNA. It was not studied further.

Combining different programs

- Use different programs, but based on different classes of algorithms (see below)
- Play with different parameters, to probe for stability. Keep everything
- Have a visualization program handy for looking at and comparing assemblies
- Usually, recurrence and stability is a good sign

Mandatory reading

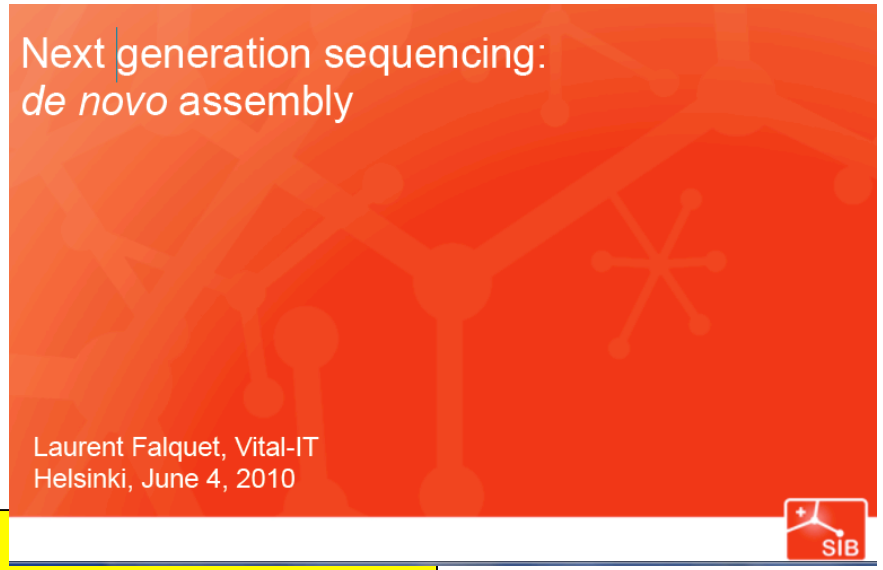
Assembly algorithms for next-generation sequencing data

Jason R. Miller *, Sergey Koren, Granger Sutton

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville MD 20850-3343, USA

Genomics 95 (2010) 315–327

Phrap: www.phrap.org



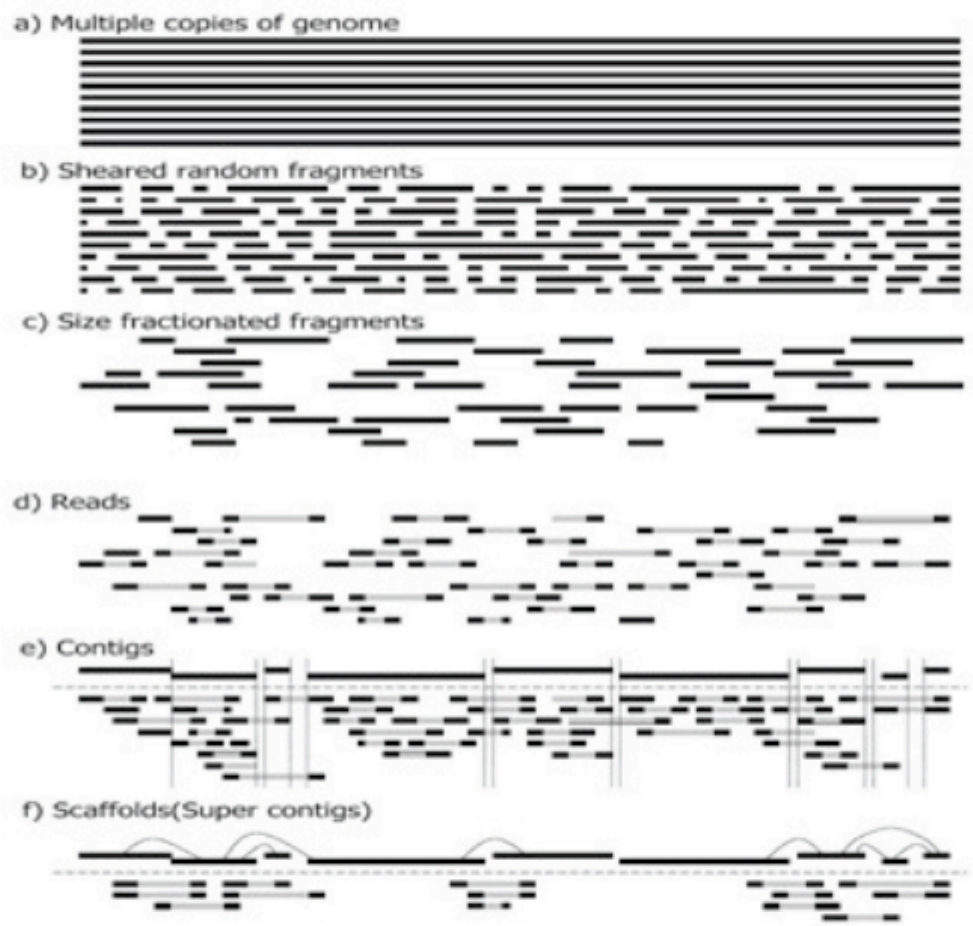
Greedy approach -> Phrap, CAP3
de Bruijn graph approach -> Velvet, Soapdenovo
Overlap/layout/consensus approach -> Newbler

Assembly: principle

You must know the various sequencing techniques,
-Experimental parameter
(Nbr reads, length)
-Cost
-Problems/limitations
-Artifacts

-And keep current!

Today:
Sanger (capillary), 454-Roche, Illumina, Solid, Ion Torrent, PacBio

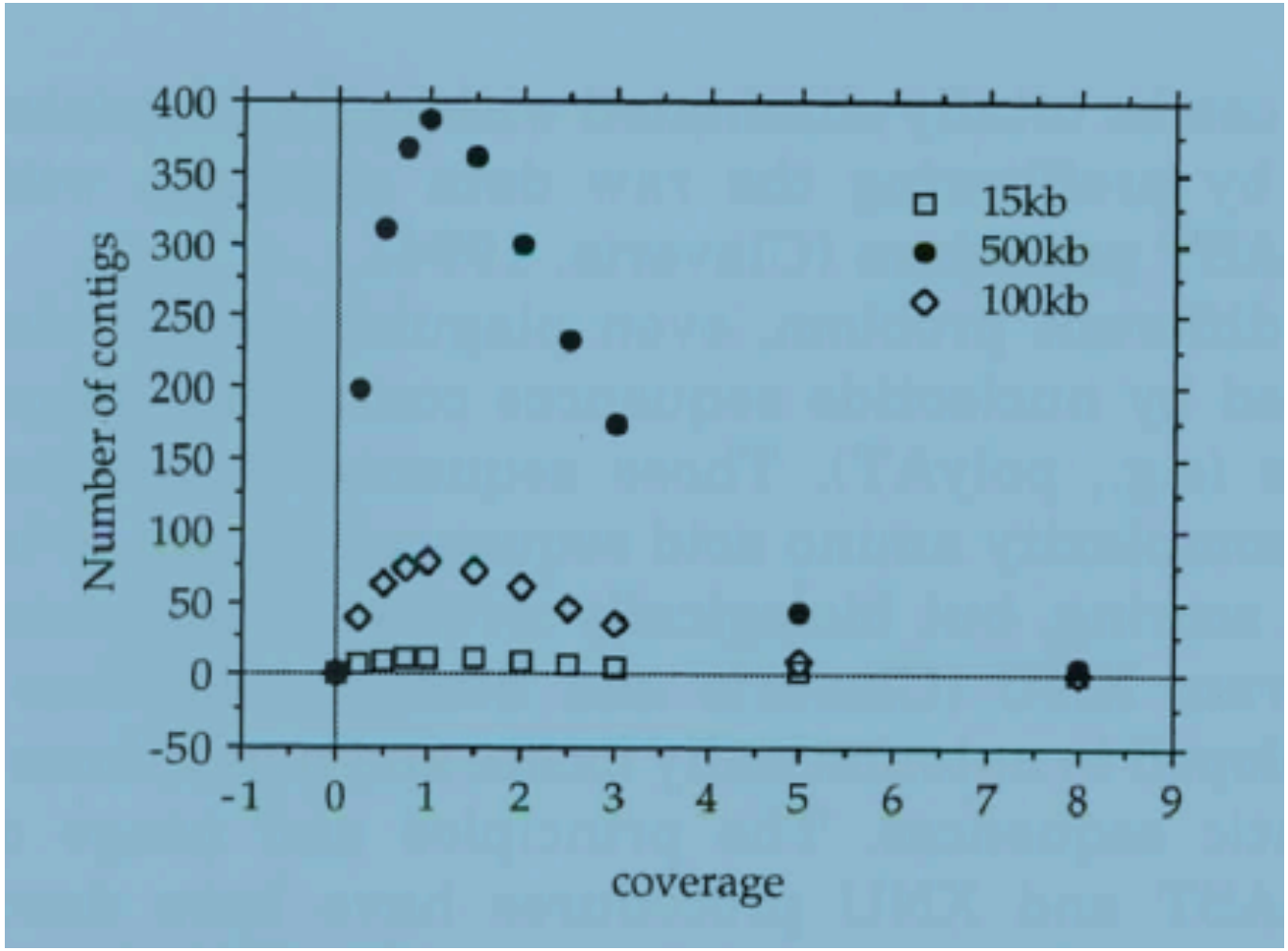


Assembly: principle

Genome size:
15 kb, 100 kb, 500 kb

Read size: 500

Overlap: 25



Compulsory reading

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER*,† AND MICHAEL S. WATERMAN‡

GENOMICS 2, 231–239 (1988)

Also:

A Streamlined Random Sequencing Strategy for Finding Coding Exons

GENOMICS 23, 575–581 (1994)

JEAN-MICHEL CLAVERIE[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894

The basic equation

$$n = c\frac{G}{R}e^{-c\left(\frac{R-O}{R}\right)}$$

n : number of remaining contigs
R: read size (bp)
G: genome size (bp) (a single piece)
O: min overlap size (bp)
c: coverage: N.R/G
N: number of reads

Exemple: Illumina sequencing of a bacteria: how much coverage is needed?

n : contig number = 1
R: read size (bp)= 100
G: genome size (bp)= One megabase
O: min overlap size (bp)= 90
c(1): coverage: N(1).R/G = ?

Exemple: Illumina sequencing of a bacteria: how much coverage is needed?

$$1 = c(1)\frac{10^6}{100}e^{-c(1)(\frac{100-90}{100})}$$

Take the $\log_{10}$:

$$0 = 2.3\ Log\big(c(1)\big) + 2.3\ 4\ - c(1)/10$$

$$23\ Log\big(c(1)\big) + 92 = \ c(1)$$

$$c(1) \approx 140$$

Exemple: Illumina sequencing of a bacteria: how much coverage is needed?

$$c(1) \approx 140$$

Thus we need 140 10^6 worth of read data,

corresponding to 1.4 10^6 reads

There is 200 M reads in one lane of Illumina HiSeq:

It is customary to use multiplexing (at least 10 bacteria/lane)

(average genome size: 4 Mb ->6 10^6 reads/genome, 3 times the Theoretical C(1) value= 420)

Exercice: how many reads to sequence a vertebrate proteome (20.000 genes) ?

Exercice: how many reads to sequence a vertebrate proteome (20.000 genes) ?

$$1 = c(1)\frac{20.10^6}{100}e^{-c(1)\left(\frac{100-90}{100}\right)}$$

$$0 = 2.3\,Log\left(c(1)\right) + 2.3\,5.3 - c(1)/10$$

$$0 = 23\,Log\left(c(1)\right) + 122 - c(1)$$

$$c(1) \approx 170$$

Exercice: how many reads to sequence a vertebrate proteome (20.000 genes)

170. 20 10^6 = 3400 10^6 worth of data

Thus: 34 Millions reads

Thus one lane worth of cDNA library
is 5.8 times more than what we need
(accounting for variable expression levels):

Transcriptome studies are cheap

Validating an assembly

- Estimating the length of a genome from the progression of the number of contigs with the number of reads used (could be a posteriori)

$$G = \frac{NR(1-\theta)}{(\ln \frac{N}{n})}$$

- Or using the n_{max} *vs* N_{read} equation

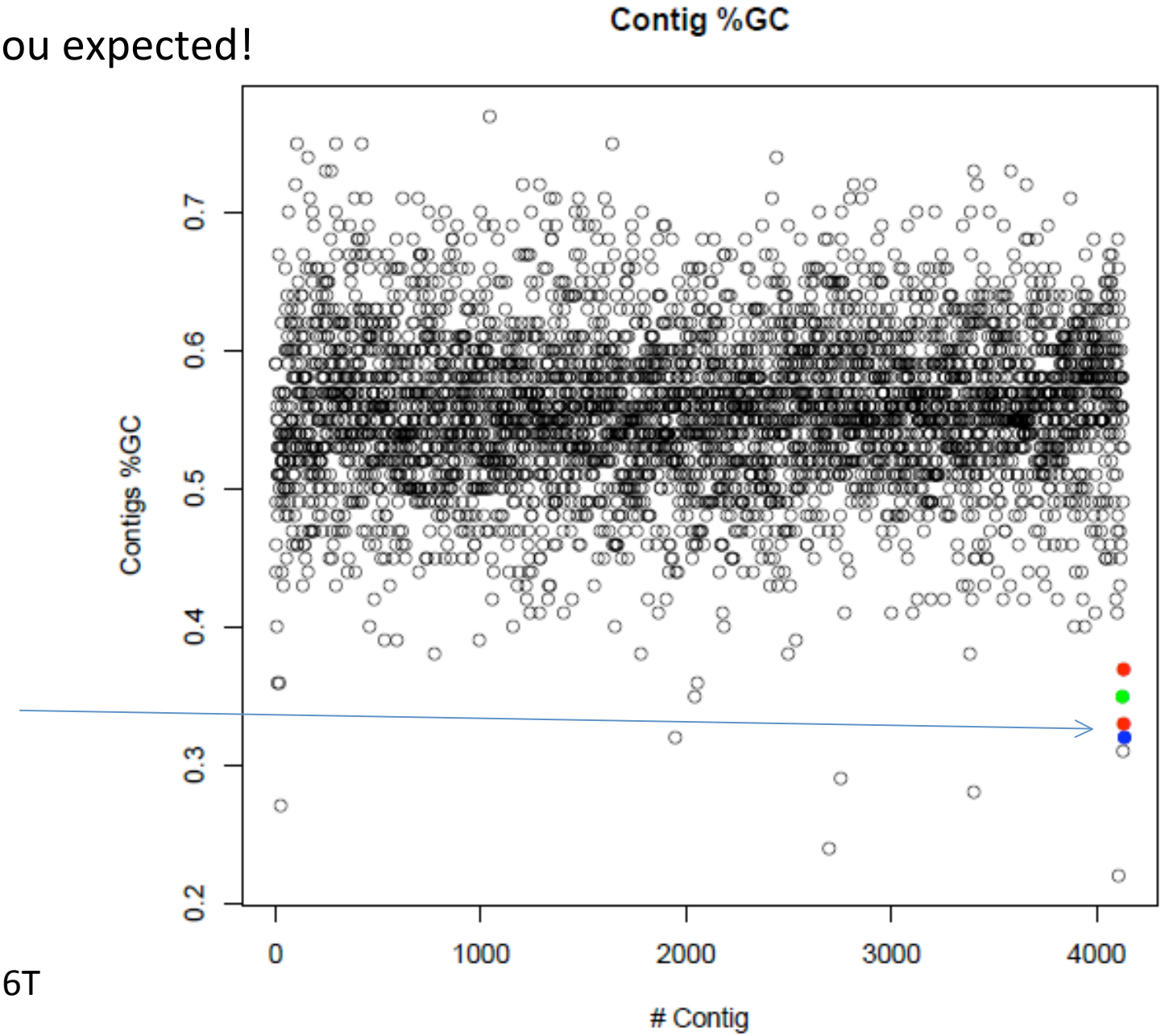
$$G = e\, R(1-\theta) n_{max}$$

Assembly: practical issues, program parameters

- Quality of read: soft vs hard « trimming »
- Overlap size (k-mers)
- Overlap: hard vs soft
- Paired-ends (or not) (beware of program « features »)
- Check the stability with random subset of the data (10% is often more than enough).

Analyse your contigs:

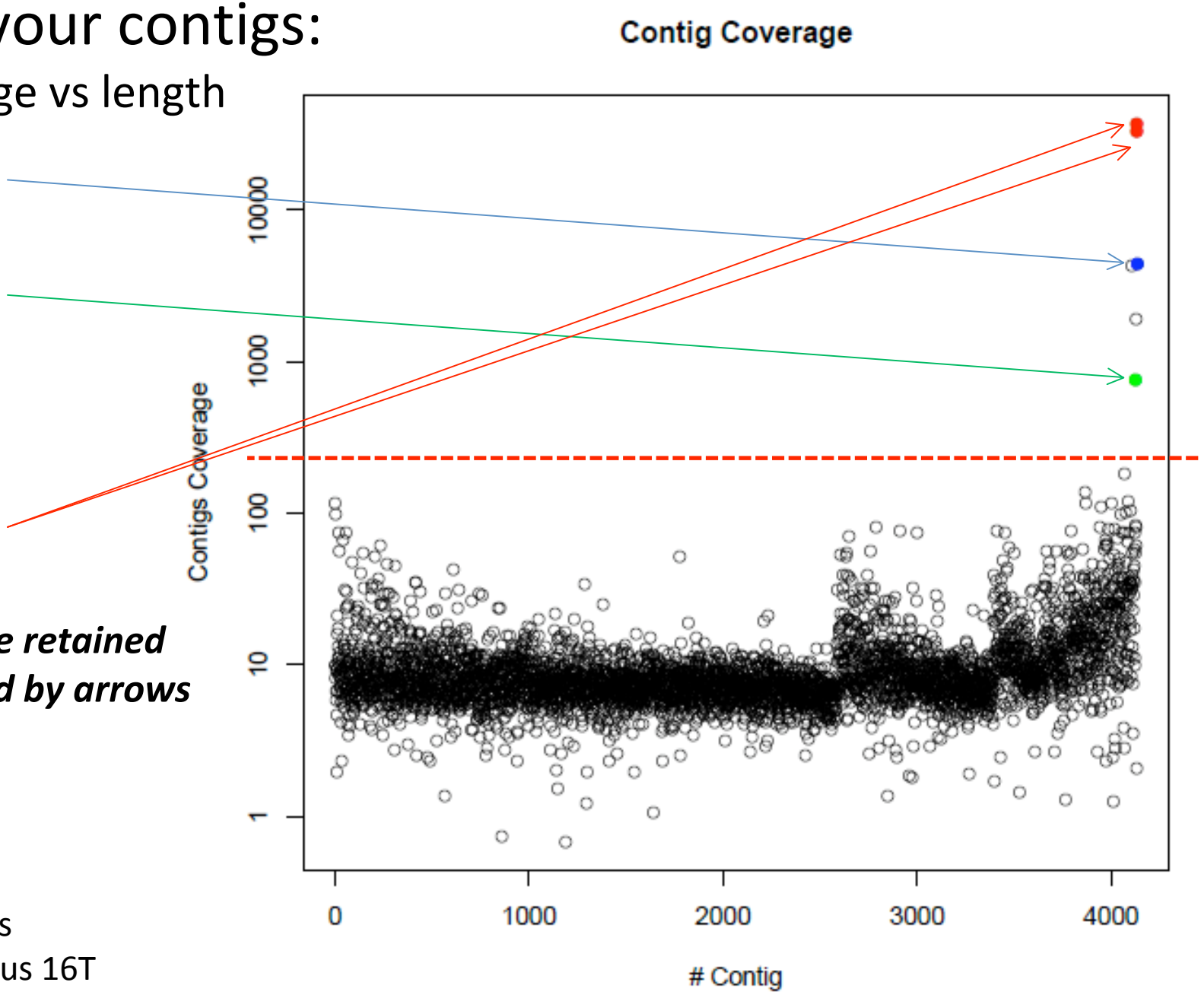
more than you expected!



Phaeocystis
globosa virus 16T

Analyse your contigs:

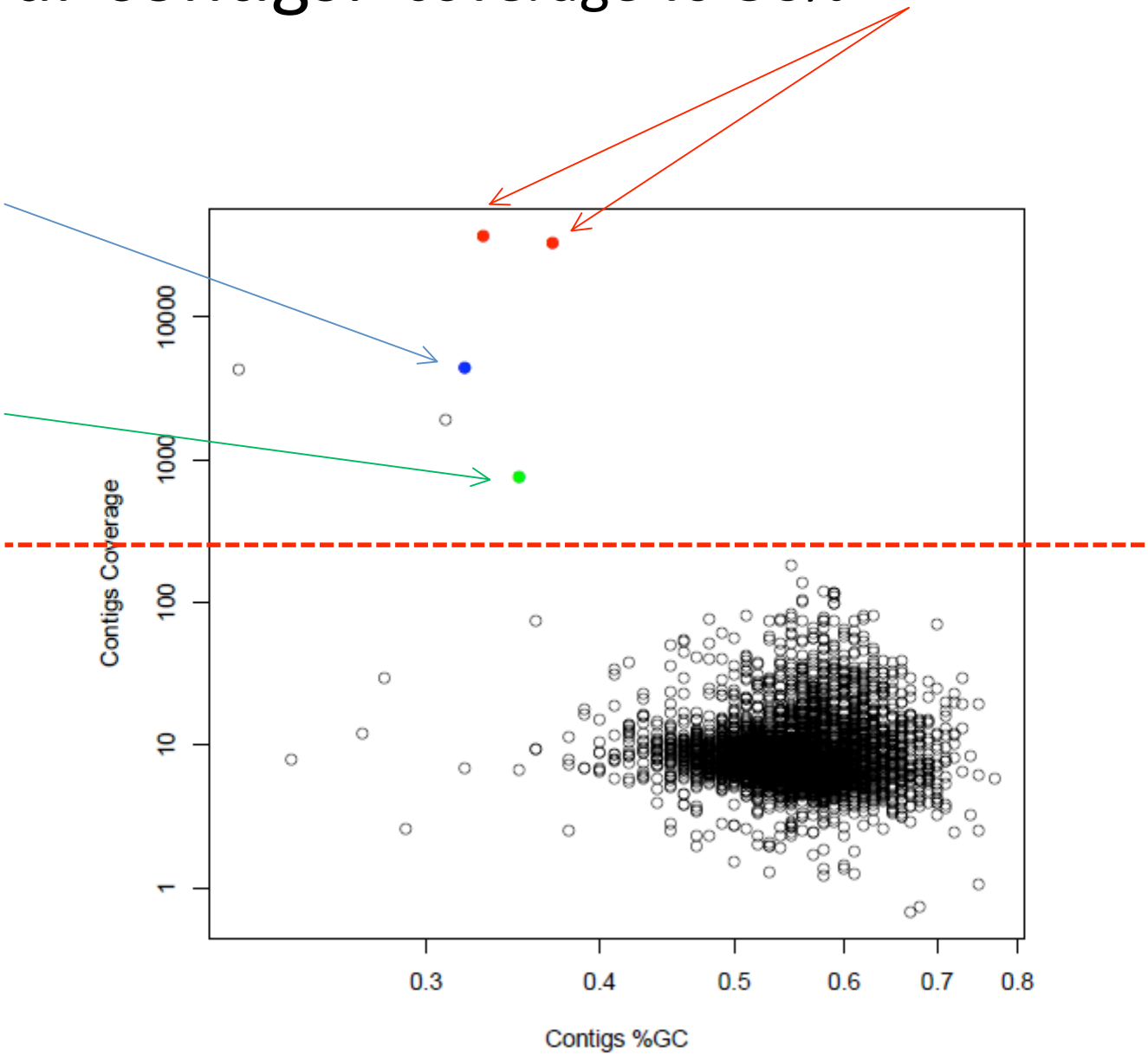
coverage vs length



Contigs to be retained are indicated by arrows

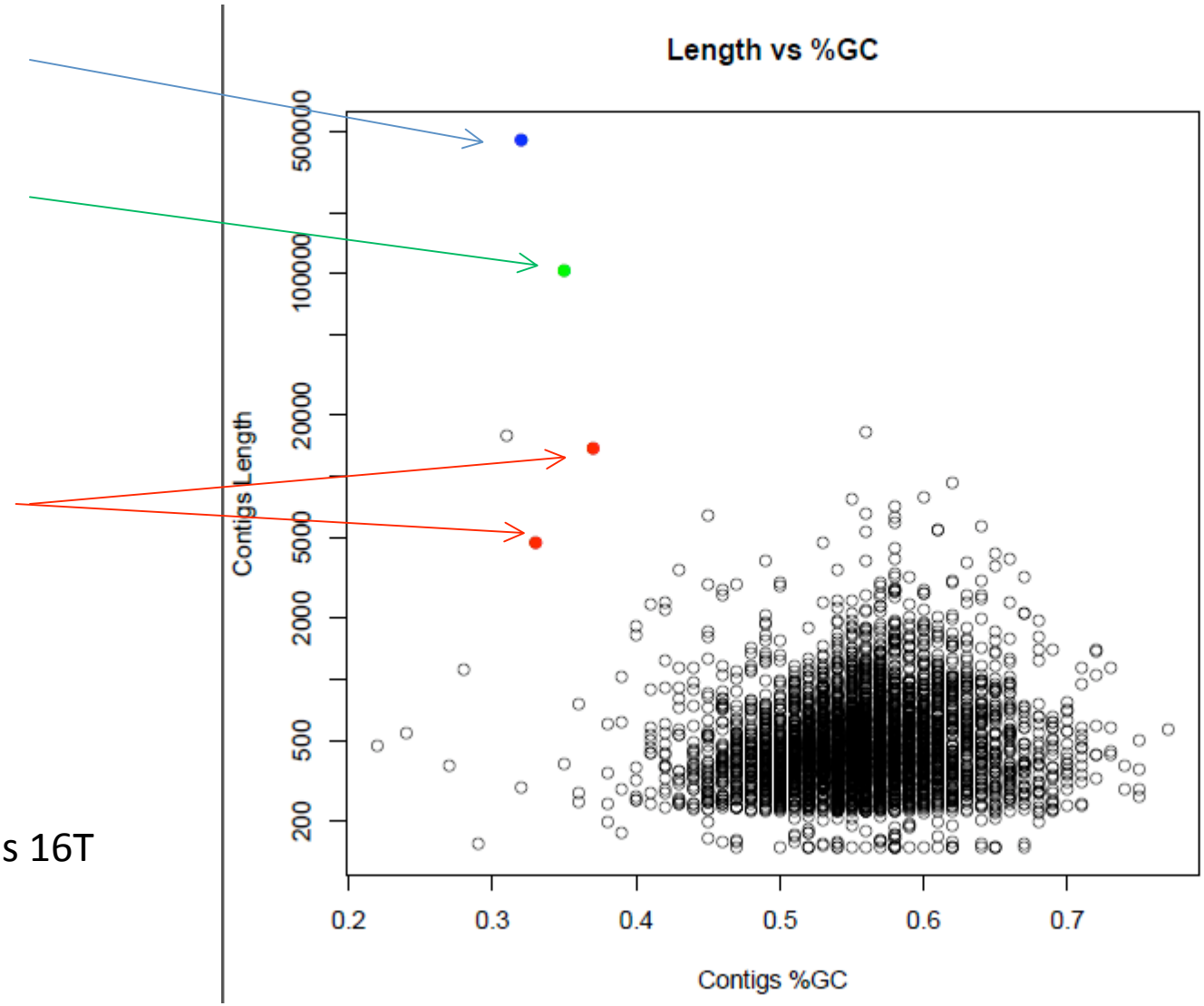
Phaeocystis globosa virus 16T

Analyse your contigs: coverage vs GC%



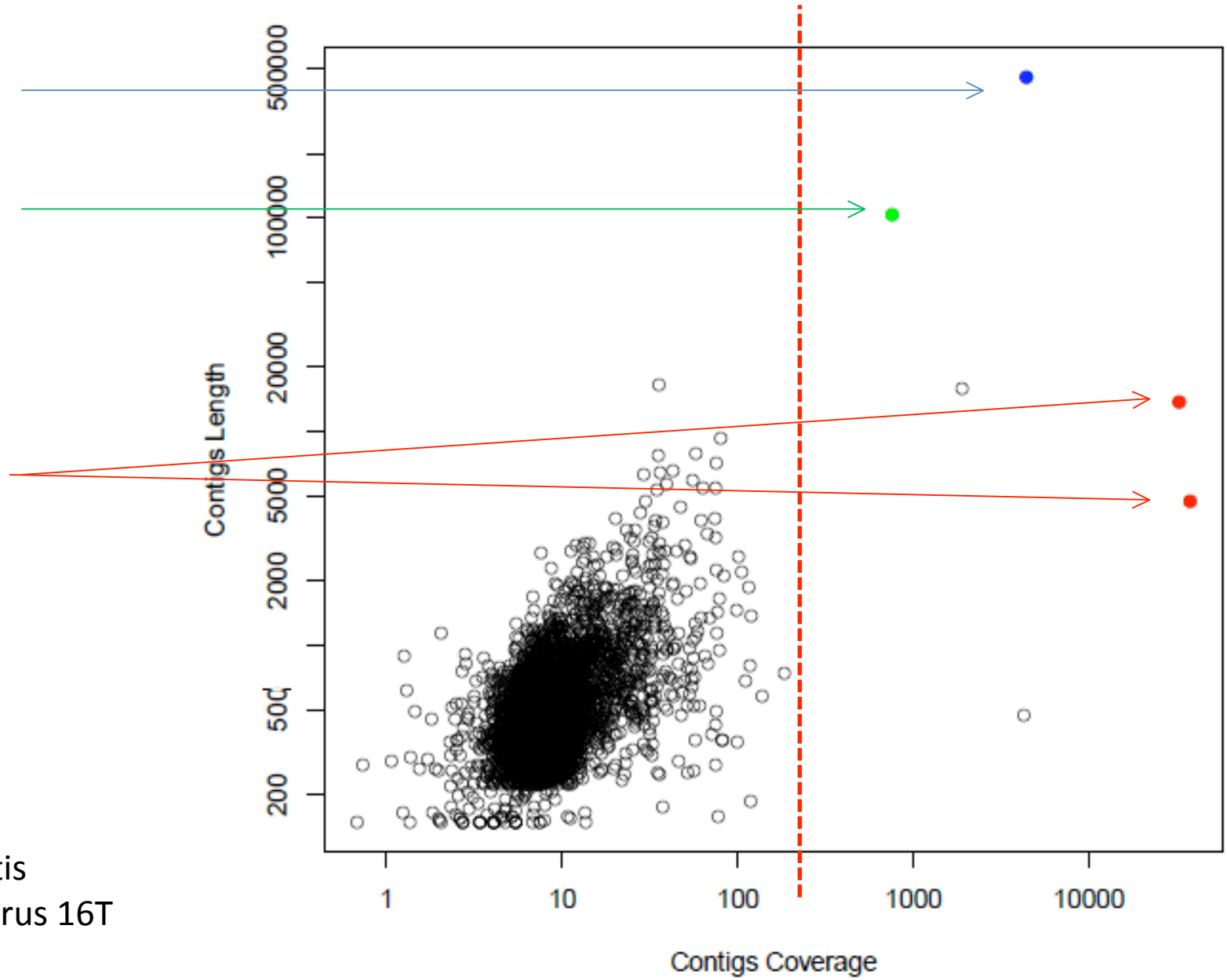
Phaeocystis
globosa virus 16T

Analyse your contigs: length vs GC%



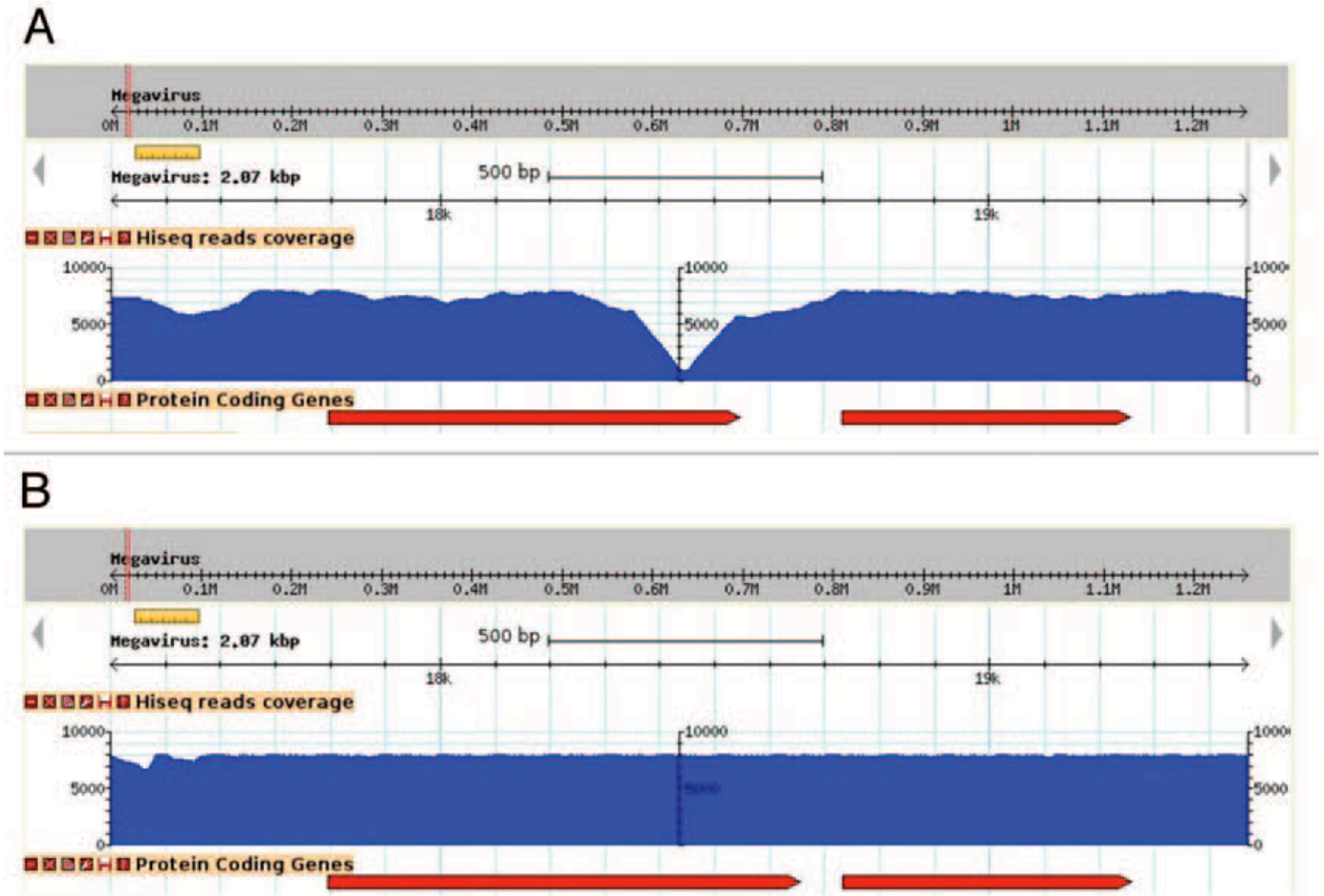
Phaeocystis
globosa virus 16T

Analyse your contigs: length vs coverage

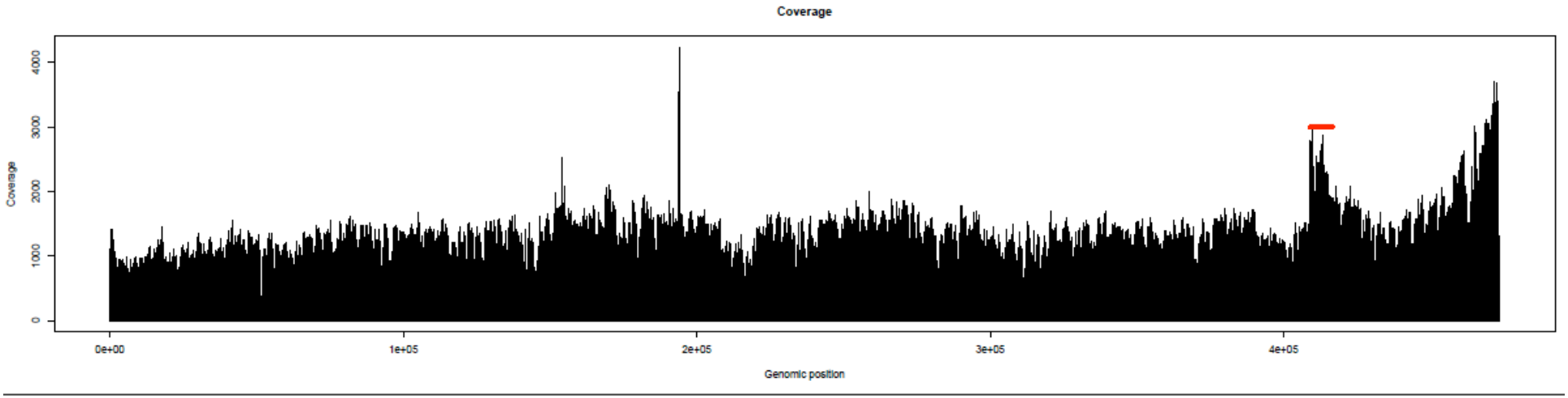


Phaeocystis
globosa virus 16T

Validation of the final assembly (Megavirus)



Validation of the final assembly: detection of repeated regions



Origins of parasitic contigs

- Impure DNA (not axenic or pure culture)
- Contamination at the sequencing level (lib prep)
  (compare the contigs built from reads from 2 # methods)
- Contamination from culture medium
- Contamination from commercial reagents
- Mixing up of tubes, multiplexing tags
- Internal control used by sequencing services
- Symbionts, prey, …, etc
- ?

Annotation: ORFing

- Orfing is MAINLY based on size!
- Orfing is possible because random long ORF (Stop to Stop) are unlikely in « regular random sequence » (A=T=G=C)
- Statistical bias (Markov models) does not work 100% (far from it!) and requires some minimal length anyway
- Orfing is possible because STOP codons are WELL respected signals
- Fortunately, proteins are long in average (250 aa)
- But some are short (ribosomal)
- Unfortunately, introns have been invented

ORFing: quick math (dont try this at home)

- Basic facts: 3 STOP codons out of 64, thus p_{stop}= 1/20, thus in a long sequence, average STOP-STOP are 20 codons in average (mean=20)
- Use the property of the Poisson distribution mean= variance
- Distribution of Stop-Stop in random seqs: Mean=20, Std. Dev= 5
- Significance at 1/1000 requires at least 3x Std -> 35 codons
- But there are **6** reading frame so 35 codons has a p value of about 1/100: 35x3= 105 amino-acids is the low limit for « significant ORFs »

ORFing: quick math, follow up

- Considering ATG-STOP makes them much more significant
  ATG-STOP are 1/60 less expected than Stop-Stop

- Unfortunately, this only work for prokaryotes and good viruses

- Other genes are fragmented:
  Spliceosomal introns (GT–AG),
  Type 2 and Type 1 self-excising introns
  inteins (in DNA – processing enzymes)
  Sequences errors -> frameshifts

- Starting by Stop-Stop « first look » is recommended when sequencing the unknown.

Our hierarchical strategy

- A+++: Genemark + DBhomolog + fonction + proteomics
- A++: Genemark + DBhomolog + function
  or Genemark + DBhomolog + proteomics
- A+: Genemark + DBhomolog
  or Genemark + proteomics
  or Genemark + homolog in close relative
- A: ORF (>40 aa) + ***relevant*** DBhomolog (i.e. ribosomal prot.)
  or (>40 aa) + homolog in close relative
- B: Genemark (>40aa)
  or ORF > 99 aa (but no overlap with above,
  no homolog whatsoever)

No overlap allowed with ORF from a higher precedence
Homolog searched using BlastP with 10^{-5} E-value

Dealing with fragmented genes

- Detected from fragmented alignment with
  - DB sequences with known function (Two ORFS, one target)
  - Homologs in relatives

- Detected by direct intein search (InteinDB, New England Biolab)

- Verify “intergenic” region using BlastX
- Impossible to find if they occur in Orfans using Bioinformatics
- Complement by transcriptome sequencing
  - Delimit gene boundaries
  - Find additional split genes
  - Locate “non coding” transcripts
  - Required for promoter analysis and motif search

Final comments

- High level publication requires Genome + Transcriptome and/or Proteome and/or experimental validations
- Exhaustive gene finding and annotation requires a full transcritome analysis. Bioinformatics-only annotation is NOT possible
- You may combine Genome closing/validation/finishing with transcriptome sequencing
- Beware of automated functional annotation:
  - Transitity error
  - implement Pfam, CDD search
  - Inspect active site, conserved residues (biblio!)
  - Look for inconsistency in the blast hit list
  - Many > “hypothetical” have a known motif!
  - Promoter/DNA motif search is rarely convincing

Back to the Introduction

- Try to learn/understand things at the conceptual level, don't be side tracked by details
- Important biological discoveries don't need statistics, in fact many basic fact are not « statistically significant »
- **LOOK** at your data: automated processing is necessary but « automated discovery » is a myth
- Try to understand the detail of the experimental protocol
- Ultimately look for surprises, not what is known
- Know the current dogma in your field of research (for eventually destroying it)
- Know your basic biochemistry and cellular biology:
  (was is anecdotal, what is central, what does not suffer exceptions)

From DNA to the best possible paper

- Having a complete genome is important
- Automated genome annotation is a myth

  (Even supervised Bioinformatics-only genome annotation is)
- Quickly assess your « genome potential »
  - New global features (ex: diploid when haploid is expected)
  - New unique features (ex: no DNA polymerase, new pathway)
  - New taxonomical feature (ex: new domain of Life)
  - Incredible **lack of novelty** (ex: unexplained pathogenicity)

Questions?